JRRD
Volume 53, Number 6, 2016
Pages 797–812

Psychometric evaluation of self-report outcome measures for prosthetic applications

Brian J. Hafner, PhD;[1*] Sara J. Morgan, PhD, CPO;[1] Robert L. Askew, PhD, MPH;[2] Rana Salem, MA[1]
[1]Department of Rehabilitation Medicine, University of Washington, Seattle, WA; [2]Department of Psychology, Stetson University, DeLand, FL

Abstract—Documentation of clinical outcomes is increasingly expected in delivery of prosthetic services and devices. However, many outcome measures suitable for use in clinical care and research have not been psychometrically tested with prosthesis users. The aim of this study was to determine test-retest reliability, mode of administration (MoA) equivalence, standard error of measurement (SEM), and minimum detectable change (MDC) of standardized, self-report instruments that assess constructs of importance to people with limb loss. Prosthesis users (N = 201) were randomly assigned to groups based on MoA (i.e., paper, electronic, or mixed mode). Participants completed two surveys 2 to 3 d apart. Instruments included the Prosthetic Limb Users Survey of Mobility, Prosthesis Evaluation Questionnaire–Mobility Subscale, Activities-Specific Balance Confidence Scale, Quality of Life in Neurological Conditions–Applied Cognition/General Concerns, Patient-Reported Outcomes Measurement Information System, and Socket Comfort Score. Intraclass correlation coefficients indicated all instruments are appropriate for group-level comparisons and select instruments are suitable for individual-level applications. Several instruments showed evidence of possible floor and ceiling effects. All were equivalent across MoAs. SEM and MDC were quantified to facilitate interpretation of outcomes and change scores. These results can enhance clinicians' and researchers' ability to select, apply, and interpret scores from instruments administered to prosthesis users.

Key words: amputation, artificial limbs, health surveys, outcome assessment (health care), outcome measures, outcomes research, prosthesis, questionnaires, rehabilitation, reproducibility of results.

INTRODUCTION

Prosthetists and other health care professionals are increasingly encouraged or required to document the effects of the care they provide using valid and reliable outcome measures [1–3]. Self-report instruments (i.e., surveys answered directly by a patient) are well suited to clinical applications because they are often brief and easy to complete. Further, information derived from self-report is often distinct and essential to understanding the effect of health interventions on the lives of those who receive them. In spite of these benefits, use of standardized outcome

Abbreviations: ABC = Activities-Specific Balance Confidence Scale, CI = confidence interval, DIF = differential item function, ICC = intraclass correlation coefficient, MDC = minimum detectable change, MGCFA = multigroup confirmatory factor analysis, MoA = mode of administration, NQ-ACGC = Quality of Life in Neurological Conditions–Applied Cognition/General Concerns, PEQ-MS = Prosthesis Evaluation Questionnaire–Mobility Subscale, PLUS-M = Prosthetic Limb Users Survey of Mobility, PROMIS = Patient-Reported Outcomes Measurement Information System, PROMIS-29 = PROMIS 29-Item Profile, SCS = Socket Comfort Score, SD = standard deviation, SEM = standard error of measurement.

***Address all correspondence to Brian J. Hafner, PhD; University of Washington, Department of Rehabilitation Medicine, 1959 NE Pacific St, Box 356490, Seattle, WA 98195; 206-685-4048. Email: bhafner@uw.edu**

http://dx.doi.org/10.1682/JRRD.2015.12.0228

measures in clinical practice remains limited [3–6]. Recommendations for instruments suited to clinical care and research involving people with lower-limb loss, similar to those that exist for other rehabilitation populations [7–8], may address barriers to outcome measure use and facilitate improved understanding of prosthetic outcomes. However, to develop formal recommendations, evidence of each instrument's performance in the population of interest (e.g., persons with lower-limb loss) is needed. Specifically, evidence of key psychometric properties (e.g., reliability, mode of administration [MoA] equivalence, measurement error, and detectable change) is required to adequately formulate recommendations for how each instrument may be applied.

Evidence of reliability (i.e., reproducibility) within the population of interest is critical for determining an instrument's utility or the applications for which it can be recommended [9]. It is generally accepted that an instrument must demonstrate test-retest reliability of 0.7 or greater to be recommended for group-level comparisons [10–15]. Group-level comparisons are important in clinical trials, observational research studies, and clinical quality-improvement programs. For applications that involve decisions about individual patients or research participants, an instrument must possess much higher (i.e., >0.9) reliability [12,15–17]. Evidence of test-retest reliability therefore becomes a key factor in distinguishing among those instruments that can be recommended for individual-level decisions and those that can be recommended for group-level comparisons.

Evidence of MoA equivalence, or performance across different forms of the same instrument, is needed to demonstrate that scores obtained from different MoAs are directly comparable [18–19]. Electronic administration via computer or tablet offers numerous benefits compared to paper surveys, including reduced respondent burden, automated and accurate scoring, and direct import into a medical or research record. Equivalence of paper and electronic MoAs would allow administrators to reap these benefits and have the flexibility to choose the format most appropriate for the respondent (e.g., paper surveys can be given to patients who may not be comfortable with technology). MoA equivalence requires that variations in scores in single-mode (e.g., paper-paper or electronic-electronic) and mixed-mode (e.g., paper-electronic) administrations be equivalent [19]. Evidence of MoA equivalence is needed to guide recommendations that may benefit from use of multiple administration methods. For example, clinics may want to give patients the option of paper surveys or computerized surveys administered on tablet computers.

Lastly, estimates of measurement error and detectable change are required to evaluate and interpret differences or changes in scores observed when using self-report instruments [15]. Estimates of measurement error, such as standard error of measurement (SEM), are used to quantify uncertainties in scores or differences in scores obtained between individuals or between groups of individuals. Estimates of detectable change, such as minimum detectable change (MDC), describe a statistical threshold (e.g., 90% or 95% confidence interval [CI]) for score differences to be considered "true" change in the context of repeated assessments (i.e., a change in outcome above and beyond that expected from measurement error) [20–21]. Estimates of change are critical in longitudinal applications, as observed changes that do not exceed an MDC may not indicate a true change in outcome and should be interpreted with caution. Estimates of measurement error and detectable change are essential to formulating recommendations based on an instrument's potential to assess changes or differences in outcomes.

The importance of the aforementioned psychometric properties cannot be overstated. If self-report outcome measures are to be used with confidence in clinical practice or research, evidence of their performance is needed to justify their selection, use, and interpretation. Few self-report instruments have been evaluated for evidence of test-retest reliability or measurement error in large samples of prosthetic limb users [1,22]. To date, none have been evaluated for MoA equivalence. Thus, there is a scarcity of evidence required to formulate use recommendations for patients or research participants with lower-limb loss. The aim of this research was to acquire the evidence needed to formulate initial recommendations for use of self-report outcome measures in prosthetic clinical care and research. Specifically, we (1) assessed test-retest reliability, (2) evaluated equivalence between paper and electronic MoAs, and (3) derived estimates of SEM and MDC for several self-report measures that are well suited to quick and efficient assessment of prosthetic outcomes. Results were also used to develop recommendations about measures most appropriate for clinical and/or research applications (e.g., measuring changes in patients over time in clinic settings or measuring differences between groups in research studies).

METHODS

Participants

Participants with lower-limb loss were recruited through the University of Washington Department of Rehabilitation Participant Pool, a national registry of individuals interested in participating in rehabilitation research. Individuals in the participant pool with limb amputation were invited to participate in the study via their preferred method of communication (i.e., mail or email). Study investigators screened interested individuals by phone, enrolled them in one of three study arms, and scheduled appointments for two survey sessions. Participants were assigned to a study arm (i.e., Arm 1a, 1b, 2, or 3) using simple randomization [23]. Eligibility criteria included (1) 18 yr of age or older; (2) lower-limb amputation between the hip and ankle; (3) amputation as the result of trauma, dysvascular complications (e.g., diabetes), infection, or tumor; (4) no other amputations (e.g., other leg or arms); (5) use of a lower-limb prosthesis to transfer or walk; (6) access to an electronic device with an Internet connection; and (7) ability to read, write, and understand English.

Study Design

We employed a three-arm randomized design (**Figure 1**) to compare scores from standardized outcome measures administered at different times and by different MoAs to participants with lower-limb loss. Each participant was scheduled a "test" and "retest" (i.e., follow-up) survey approximately 2 to 3 d apart. A minimum period between surveys of 2 d was targeted to mitigate the potential for recall bias [12]. Similarly, the maximum duration between surveys was targeted to minimize natural changes in the selected outcomes (e.g., mobility, physical function, balance). At retest, participants were asked to indicate whether they had experienced any changes in health status since the test survey. Those who indicated their health had changed were excluded from the final data set. Participants in Arm 1 received one paper survey and one electronic survey. The order of the surveys (i.e., paper-electronic [1a] or electronic-paper [1b]) was assigned randomly, as recommended by Coons et al. [24]. Participants in Arm 2 were administered two paper surveys, and participants in Arm 3 were administered two electronic surveys.

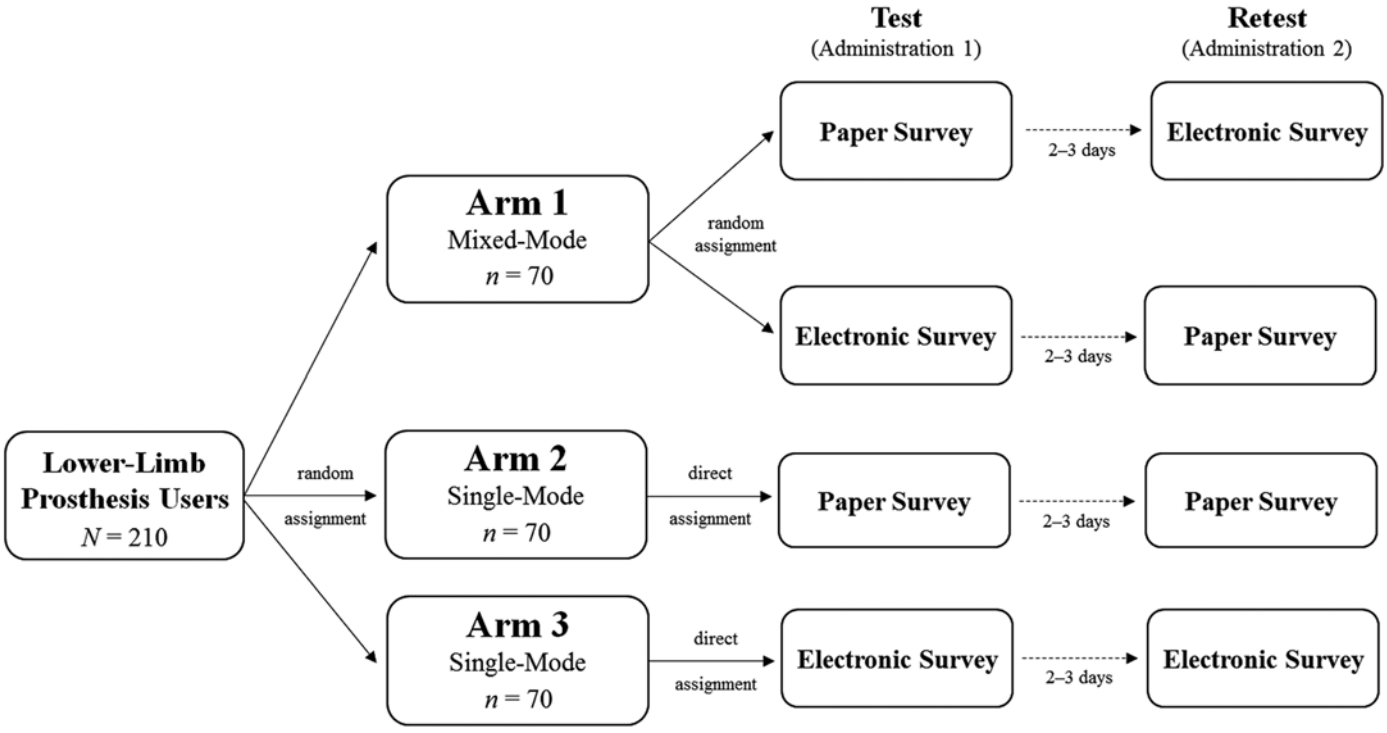

Figure 1.
Study design overview. Participants were assigned to arms based on survey mode of administration and completed surveys twice within 2–3 d. Sample sizes reflect the number of participants targeted, based on an a priori sample size estimation.

Sample Size

Minimum sample size was calculated using the methodology outlined by Walter et al. [25], using $\alpha = 0.05$, $\beta = 0.20$, $\rho_0 = 0.50$, $\rho_1 = 0.70$, and $n = 2$ assessments. The lower bound of $\rho_0 = 0.50$ was selected because the suitability of measures with this level of reliability is questionable, even for group-level comparisons. The sample size target was increased from 63 to 70 per arm (from $N = 189$ to $N = 210$ across all arms) to account for possible attrition of study participants or changes in health status between the test and retest surveys.

Surveys

Paper surveys were printed on standard letter paper, sealed in individual envelopes, and mailed to participants with instructions and a self-addressed return envelope. Instructions indicated that participants were to open the envelope only at the time of their scheduled appointment. Electronic surveys were created and administered using the Assessment Center (Northwestern University, Chicago, Illinois) [26]. Electronic questions were identical to paper questions, but formatting differed slightly to facilitate item-level computerized administration (e.g., response options were presented vertically on computer and horizontally on paper). Electronic surveys were uniquely coded to each participant and sent via an email link with instructions similar to the paper surveys (i.e., that the survey was not to be started until the time of the scheduled appointment). Paper survey responses were double-entered by research staff to minimize data-entry errors [27]. Electronic responses were exported directly from the Assessment Center for analysis. Responses from all participants were screened for missing and/or potentially invalid responses. Participants were contacted to clarify responses as needed.

Measures

Test and retest surveys included questions on demographics and participant characteristics in addition to the standardized self-report measures described subsequently. An ad hoc question was also included in the retest survey to solicit any changes in respondents' health between survey time points. In addition, participants were asked to record the time that they began and ended the paper surveys to calculate the time required to complete each survey. Time to complete electronic surveys was recorded by the Assessment Center administration system.

Demographics and Participant Characteristics

Demographic information (e.g., sex, race/ethnicity, employment status) and participant characteristics (e.g., height, weight) were collected to describe the study sample. In addition, participants answered questions related to their amputation (e.g., date, cause, and level of amputation) and health (e.g., presence of comorbidities) to characterize their general health.

Outcome Measures

Six self-reported outcome measures suited to prosthetic applications were assessed in this study. The Prosthetic Limb Users Survey of Mobility (PLUS-M) is an item bank developed to measure perceived mobility in people with lower-limb amputation [28–30]. The PLUS-M 12- and 7-item short forms were both administered in this study. Additionally, the PLUS-M computerized adaptive test was administered to participants in Arm 3 (i.e., electronic-electronic). The Prosthesis Evaluation Questionnaire–Mobility Subscale (PEQ-MS) is a 12-item self-report measure assessing the ability to perform mobility tasks while using a lower-limb prosthesis [31]. The Activities-Specific Balance Confidence Scale (ABC) is a 16-item instrument that measures respondents' confidence in performing basic ambulatory activities [32]. Recent Rasch analyses of the PEQ-MS and ABC resulted in similar recommendations to reduce the instruments' original visual analog scale (PEQ-MS) and 0–100 (ABC) response options to 5-point ordinal scales [33–34]. These recommended modifications were incorporated into the instruments administered in this study [33–34]. The Quality of Life in Neurological Conditions–Applied Cognition/General Concerns (NQ-ACGC) v1.0 is an item bank developed to measure general cognitive abilities, including memory, attention, and decision-making [35]. The 8-item NQ-ACGC short form was administered in this study. The Patient-Reported Outcomes Measurement Information System (PROMIS) is a compilation of self-report instruments that measures eight symptom and quality-of-life constructs across patient populations: Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Social Role Satisfaction, Pain Interference, and Pain Intensity [36–37]. The PROMIS 29-Item Profile (PROMIS-29) was administered to participants in this study. The Socket Comfort Score (SCS) is a one-item measure of prosthetic socket comfort [38]. Participants' scores were calculated according to developers' instructions and used to evaluate test-retest reliability, MoA equivalence, SEM, and MDC. The ABC and PEQ-MS are scored from 0 to 4 (i.e., average score of all items), and the SCS is scored from 0 to 10 (i.e., score of the

single SCS item). PLUS-M, PROMIS-29 (all domains except PROMIS pain intensity), and NQ-ACGC are scored on a T-score metric, which has a mean of 50 and standard deviation (SD) of 10 [39]. PROMIS pain intensity is scored 0 to 10.

Statistical Analysis

Differences in participant demographics (e.g., sex, race/ethnicity, employment status, income, education, Veteran status, amputation level, amputation etiology) by study arm were assessed using chi-square or a Fisher exact test. Descriptive statistics (e.g., mean, SD) were calculated for participants' scores at both time points (i.e., test and retest). The distribution of each measure was evaluated for problematic departures from normality using traditional statistical tests and histogram inspections [40–41]. Mixed-effects linear regression modeling was employed to test differences in mean scores by MoA and time (i.e., test-retest) because of its recognized advantages (e.g., flexibility and robustness) over traditional analysis of variance analyses. Test-retest reliability was evaluated using the intraclass correlation coefficient (ICC) model 3,1 for individual scores using a fixed effect for time (i.e., test-retest) and a random effect for individuals [42–43]. CIs for ICCs were derived using the *F*-distribution [44]. MoA equivalence was similarly evaluated by tests of statistically significant differences using an *F*-distribution [45–46]. The a priori alpha level ($\alpha = 0.05$) was adjusted for multiple comparisons using a Bonferroni correction [47]. Given that statistically significant differences in ICCs may not always affect recommendations regarding a measure's suitability for clinical or research applications, we also subjectively assessed the ICCs across modes based on the recommended thresholds (i.e., 0.7 for group-level comparisons and 0.9 for intra-individual comparisons) [10–17]. Accordingly, when ICCs for each outcome measure (across MoAs) were similar (i.e., >0.9, 0.7–0.9, or <0.7), global test-retest reliability estimates were computed for participants across all modes. SEM and MDC estimates were derived using established algebraic transformations based on calculated ICCs and *z*-scores for the 90 percent and 95 percent CIs [43].

RESULTS

In total, 219 participants completed all study procedures (**Figure 2**). Given the delay in participants returning surveys, we recruited participants until we had sufficient numbers of returned surveys; as additional surveys were returned after that time, we ended up with a larger sample than expected. Eighteen participants reported changes in health between the test and retest time points and were excluded from the final data set to avoid biasing reliability estimates. A variety of changes were reported, ranging from temporary socket discomfort to hospitalization. There were no significant differences between the participants who reported a change in health status over the test-retest period and those who were included in the final data set. Similarly, there were no significant differences among study arms in terms of participants' sex, race/ethnicity, employment status, income, Veteran status, education, amputation level, amputation etiology, age, age at amputation, time since amputation, or average time of prosthesis use per day. In the final data set, participants' ($N = 201$, **Table 1**) mean ± SD age at the time of the survey was 60.2 ± 11.4 yr, age at the time of their amputation was 41.8 ± 17.3 yr, and time since amputation was 18.4 ± 17.2 yr. Participants reported wearing their prosthesis a mean ± SD of 13.4 ± 3.8 h per day. Retest surveys were taken a mean ± SD of 48.9 ± 5.2 h after the test survey, and the average time to complete the test and retest surveys was 12.3 ± 7.8 min and 10.0 ± 5.8 min, respectively. Paper surveys, on average, took longer to complete (13.8 min) than electronic surveys (8.3 min).

Observed score distributions were approximately normal for PLUS-M and PROMIS Physical Function, Fatigue, and Sleep Disturbance. Evidence of potential floor effects were observed for PROMIS Depression, Anxiety, Pain Interference, and Pain Intensity (42%, 34%, 28%, and 12% of respondents scored the minimum score on each instrument, respectively). Similarly, potential ceiling effects were observed for the SCS, PROMIS Physical Function, PROMIS Satisfaction with Social Roles, and NQ-ACGC (14%, 14%, 16%, and 17% of respondents scored the maximum score on each instrument, respectively). No evidence of floor or ceiling effects was present for the ABC, PEQ-MS, or PLUS-M. No statistically significant differences in mean scores (i.e., retest score − test score) were present between MoA groups (**Appendix 1**). Statistically significant effects of time ($p < 0.05$) were observed for five PROMIS measures (i.e., Anxiety, Depression, Fatigue, Pain Intensity, and Sleep Disturbance) and the NQ-ACGC. However, differences between test and retest scores were negligible (i.e., −1.9 to 0.2) and below minimal important

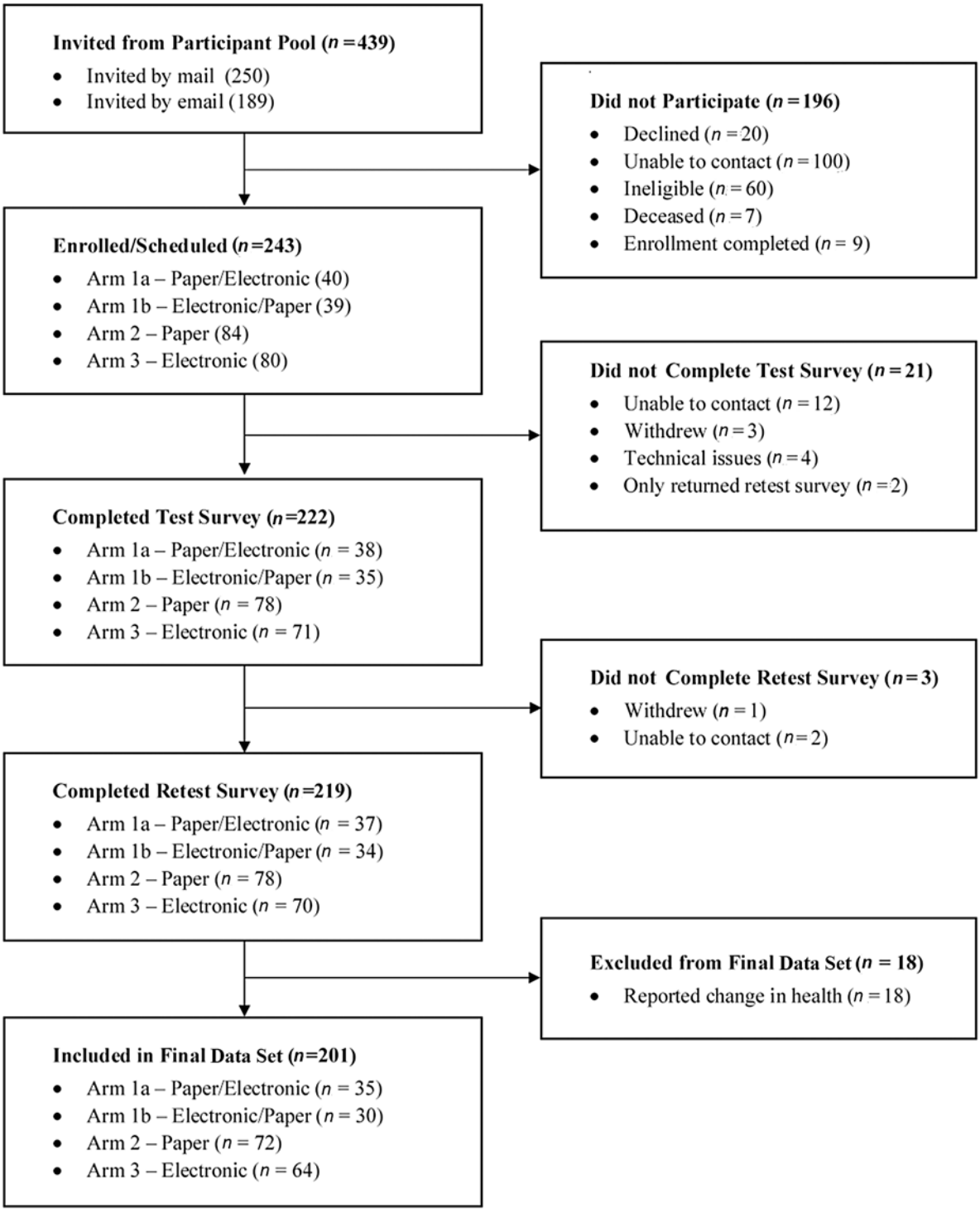

Figure 2.
Study flow diagram.

difference estimates (i.e., 2.5 to 6.0) reported for PROMIS measures in other clinical populations [48]. No statistically significant time by MoA interactions were observed.

Table 1.
Participant demographics by arm. Data presented as number (percent).

Characteristic	Arm 1 (Mixed-mode) *n* = 65	Arm 2 (Paper-only) *n* = 72	Arm 3 (Electronic-only) *n* = 64	All Arms *N* = 201
Sex*				
Male	37 (56.9)	55 (76.4)	43 (67.2)	135 (67.2)
Female	28 (43.1)	17 (23.6)	21 (32.8)	66 (32.8)
Race/Ethnicity				
Non-Hispanic White	57 (87.7)	66 (91.7)	60 (93.8)	183 (91.0)
Non-Hispanic Black	3 (4.6)	4 (5.6)	2 (3.1)	9 (4.5)
Other	5 (7.7)	2 (2.8)	2 (3.1)	9 (4.5)
Employment Status				
Employed	24 (36.9)	23 (31.9)	26 (40.6)	73 (36.3)
Homemaker	2 (3.1)	1 (1.4)	2 (3.1)	5 (2.5)
Retired	20 (30.8)	25 (34.7)	24 (37.5)	69 (34.3)
On Disability	16 (24.6)	18 (25.0)	10 (15.6)	44 (21.9)
Unemployed	1 (1.5)	5 (6.9)	2 (3.1)	8 (4.0)
Student	2 (3.1)	0 (0.0)	0 (0.0)	2 (1.0)
Individual Income				
<$25,000	30 (46.2)	21 (29.2)	19 (29.7)	70 (34.8)
$25,000–$39,999	8 (12.3)	13 (18.1)	12 (18.8)	33 (16.4)
$40,000–$54,999	5 (7.7)	8 (11.1)	6 (9.4)	19 (9.5)
$55,000–$69,999	9 (13.8)	7 (9.7)	8 (12.5)	24 (11.9)
$70,000–$84,999	3 (4.6)	6 (8.3)	5 (7.8)	14 (7.0)
$85,000–$99,999	3 (4.6)	3 (4.2)	5 (7.8)	11 (5.5)
$100,000+	6 (9.2)	13 (18.1)	7 (10.9)	26 (12.9)
Not Reported	1 (1.5)	1 (1.4)	2 (3.1)	4 (2.0)
Veteran Status				
Not a Veteran	50 (76.9)	51 (70.8)	48 (75.0)	149 (74.1)
Active/Veteran	14 (21.5)	21 (29.2)	16 (25.0)	51 (25.4)
Not Reported	1 (1.5)	0 (0.0)	0 (0.0)	1 (0.5)
Education				
High School Graduate or Less	7 (10.8)	12 (16.7)	6 (9.4)	25 (12.4)
Some College or Tech School	21 (32.3)	31 (43.1)	20 (31.3)	72 (35.8)
College Graduate	20 (30.8)	17 (23.6)	16 (25.0)	53 (26.4)
Advanced Degree	17 (26.2)	12 (16.7)	22 (34.4)	51 (25.4)
Amputation Level				
Transfemoral	26 (40.0)	24 (33.3)	20 (31.3)	70 (34.8)
Transtibial	39 (60.0)	48 (66.7)	44 (68.8)	131 (65.2)
Amputation Etiology				
Dysvascular	17 (26.2)	13 (18.1)	16 (25.0)	46 (22.9)
Trauma	36 (55.4)	49 (68.1)	36 (56.3)	121 (60.2)
Infection	9 (13.8)	8 (11.1)	8 (12.5)	25 (12.4)
Tumor	3 (4.6)	1 (1.4)	4 (6.3)	8 (4.0)
Congenital	0 (0.0)	1 (1.4)	0 (0.0)	1 (0.5)

Note: There were no significant differences in demographic characteristics between study arms ($p > 0.05$). Percentages may not add to 100 due to rounding.
*The question posed specifically stated, "Please indicate your gender."

Test-Retest Reliability

Test-rest reliability ICCs varied by instrument (**Appendix 2**). PROMIS and Neuro-QoL measures exhibited ICCs between 0.7 and 0.9, indicating they are appropriate for group-level comparisons, irrespective of MoA. ICCs for PLUS-M, PEQ-MS, and ABC were >0.9, indicating they are appropriate for individual-level monitoring and decision-making. SCS test-retest ICCs ranged from 0.63 to 0.79, depending on MoA, indicating that the SCS is appropriate for group-level comparisons, but only when administered in a single mode (e.g., either paper only or electronic only).

Mode of Administration Equivalence

ICCs by MoA varied slightly by measure (**Appendix 2**) but were generally consistent relative to the established reliability thresholds (i.e., ≥0.9, 0.7–0.9, or ≤0.7). No significant differences in ICCs were observed by MoA for any of the measures, with the exception of PEQ-MS (p = 0.04). However, all PEQ-MS ICCs exceeded the 0.9 threshold recommended for individual-level applications.

Standard Error of Measurement and Minimum Detectable Change

As most measures were determined to be equivalent across MoAs or were consistently within ranges established by the reliability thresholds, SEM and MDC estimates were derived from combined ICCs (**Table 2**). For the SCS, SEM and MDC were derived for each mode, given that instrument's low (and variable) reliability by MoA. PLUS-M showed the lowest MDC (4.50) for all instruments scored using the T-metric (i.e., NQ-ACGC, PLUS-M, and PROMIS instruments). MDC estimates for measures that used an average score (i.e., ABC and PEQ-MS) were comparable (i.e., 0.49 and 0.55, respectively). Interestingly, MDC for the SCS (2.73) was larger than the PROMIS Pain Intensity scale (1.97), although both instruments are scored similarly (i.e., 0–10 scale).

Table 2.
Test-retest reliability, standard error of measurement (SEM), and minimum detectable change (MDC) of self-report instruments in people with lower-limb loss.

Measure	ICC	95% LB	95% UB	SEM	MDC (90% CI)	MDC (95% CI)
ABC	0.95	0.93	0.96	0.21	0.49	0.58
NQ-ACGC	0.88	0.85	0.91	2.87	6.67	7.94
PEQ-MS	0.92	0.90	0.94	0.24	0.55	0.65
PLUS-M CAT	0.92	0.87	0.95	2.79	6.42	7.65
12-Item Short Form	0.96	0.95	0.97	1.93	4.50	5.36
7-Item Short Form	0.95	0.94	0.96	2.02	4.69	5.59
PROMIS						
Anxiety	0.86	0.82	0.89	3.36	7.81	9.31
Depression	0.88	0.85	0.91	2.89	6.71	8.00
Fatigue	0.84	0.80	0.88	3.33	7.74	9.22
Pain Intensity	0.87	0.84	0.90	0.85	1.97	2.35
Pain Interference	0.82	0.77	0.86	3.66	8.51	10.14
Physical Function	0.88	0.85	0.91	2.64	6.13	7.31
Sleep Disturbance	0.85	0.81	0.89	3.27	7.61	9.07
Social Role Satisfaction	0.79	0.73	0.84	4.10	9.53	11.36
SCS						
All Modes	0.74	0.67	0.80	1.18	2.73	3.26
Mixed Mode	0.63	0.45	0.75	1.30	3.03	3.61
Paper Only	0.77	0.66	0.85	1.21	2.82	3.36
Electronic Only	0.79	0.67	0.86	0.99	2.31	2.75

Note: Reliability, SEM, and MDC are presented separately by mode of administration when differences were observed.
ABC = Activities-Specific Balance Confidence Scale, CAT = computerized adaptive test, CI = confidence interval, ICC = intraclass correlation coefficient, LB = lower bound, NQ-ACGC = Quality of Life in Neurological Conditions–Applied Cognition/General Concerns, PEQ-MS = Prosthesis Evaluation Questionnaire–Mobility Subscale, PLUS-M = Prosthetic Limb Users Survey of Mobility, PROMIS = Patient-Reported Outcomes Measurement Information System, SCS = Socket Comfort Score, UB = upper bound.

DISCUSSION

The aim of this study was to evaluate key psychometric properties for six self-report measures suitable for use in prosthetic clinical care and research. The estimates of reliability, MoA equivalence, SEM, and MDC provided here can help clinicians and researchers select, use, and interpret information provided by the studied self-report outcome measures. To our knowledge, this study is the first to assess reliability, MoA equivalence, SEM, and MDC of these instruments in people with lower-limb loss. Our sample ($N = 201$) was large relative to similar studies that assessed reliability of self-report instruments in people with lower-limb loss [49–50] and exceeded the minimum threshold (i.e., $N = 100$) recommended by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criterion for an "excellent" rating in studies of instrument reliability [51]. Demographics of participants included in this study were similar to those reported in a large national survey of 935 people with limb loss [52]. Sex and ethnicity in our study sample (i.e., 67% male and 91% non-Hispanic white) were nearly identical to those in the large national sample. Participants here were slightly older (i.e., 60 yr), and there was a larger proportion of participants with traumatic amputation (i.e., 60%) compared to the prior study (i.e., 50 yr and 39%, respectively). However, the overall similarities between samples suggest the results obtained here can be well generalized to people with lower-limb loss.

Test-Retest Reliability

Test-retest reliability provides critical information about instruments' stability of measurement when respondents are not experiencing change, and estimates of reliability obtained in these situations indicate applications for which the instrument may be appropriate. Measures with test-retest reliability estimates <0.7 have an unacceptable amount of error variance (e.g., intra-individual variation, measurement error) and thus are typically not appropriate for use in clinical care or research. Measures with test-retest reliability estimates >0.7 have an acceptable level of error variance for assessment of differences between groups [10–15]. However, when the goal of measurement is to assess true change within an individual over time, very low error variance is acceptable. Thus, a minimum test-retest reliability estimate of 0.9 has been suggested as the cutoff for intra-individual measurement [15–17].

All measures assessed in this study, with the exception of SCS, were found to have test-retest reliability ICCs acceptable for group-level comparisons (i.e., ICC ≥ 0.7), irrespective of MoA. The SCS was found to have test-retest reliability ICCs >0.7 when administered in either paper-only or electronic-only versions and thus can be recommended for group-level comparisons only when a single administration method is used. Three measures (i.e., PLUS-M, PEQ-MS, and ABC) were estimated to have test-retest reliability ICCs suitable for use in individual-level comparisons (i.e., ICC ≥ 0.9). Reliability estimates in this study for the PEQ-MS and the ABC are slightly higher than those published for the 7-level response scale version of the PEQ-MS (ICC = 0.85) and the 101-level response scale version of the ABC (ICC = 0.91), suggesting that the 5-level response scales recommended in subsequent studies (and used in the present study) may provide greater stability than the original versions of these instruments [33–34,49,53].

Mode of Administration Equivalence

MoA equivalence is essential if data are collected via different methods (i.e., paper surveys, electronic surveys) and scores are to be compared or aggregated across modes (e.g., if a practitioner were to administer a paper survey in clinic and then send a patient an email survey at follow-up). Recent meta-analyses have concluded equivalence of paper and electronic self-report instruments in both nondisabled individuals and those with a variety of medical conditions [19,54–55]. However, none of the studies included in the meta-analyses targeted (or, to our knowledge, included) people with lower-limb loss. This study addresses this gap and contributes evidence to the body of knowledge regarding MoA equivalence.

Results of our study provided evidence of statistical MoA equivalence for five of the six measures, suggesting that paper and electronic forms of these measures are directly comparable. The PEQ-MS was not statistically equivalent across modes. However, the test-retest ICCs for this instrument were high and similar across modes (i.e., all ICCs ≥ 0.9), suggesting reliability of the PEQ-MS does not meaningfully differ across modes. Further, although the statistical analysis for the SCS demonstrated equivalence between modes, mixed-mode administration of this measure resulted in test-retest ICCs below the cutoff for group-level comparisons (i.e., ICC = 0.63) and within-mode administration resulted in ICCs >0.7 (i.e., paper-only ICC = 0.77, electronic-only ICC = 0.79).

Because mixing MoAs appears to adversely affect reliability of the SCS, we recommend it be administered using only a single method (i.e., paper-only or electronic-only) across all individuals whose scores are to be combined or compared. Our results are largely consistent with prior findings [19,54–55] and indicate that paper and electronic surveys can generally be used interchangeably in people with lower-limb loss, with the notable exception of the SCS. The format in which the SCS was presented to respondents in the paper and electronic surveys (i.e., 11-point horizontal ordered response scale and 11-point drop-down menu, respectively) may have affected the cross-modal reliability of the instrument. Response options for other measures were generally more similar across modes (i.e., ABC, PEQ-MS, PLUS-M, and PROMIS all used horizontal 5-point ordered response scales in the paper mode and vertical 5-point ordered response scales in the electronic mode) and may have improved stability across modes, compared to the enumerated scale used by the SCS [56]. However, as the SCS and PROMIS Pain Intensity instruments are constructed and administered similarly, the disparate MoA equivalency results found between these two instruments were unexpected. Further research is needed to ascertain the source of mixed-mode measurement error (or variation) with the SCS.

Measurement Error and Detectable Change

Estimates of measurement error (and detectable change) can be used to evaluate individuals' scores with respect to threshold (i.e., cutoff) values or previous measurements [43]. They can also help in determining sample size estimates in group research [12]. SEM and MDC values obtained in this study were derived in a manner similar to a recent study by Resnik and Borgia that examined reliability of performance-based and self-report instruments in 44 people with lower-limb loss [49]. Direct comparison of SEM and MDC values is difficult because instruments differed between studies. However, the values we obtained were similar to those they derived, when evaluated as a percentage of each instrument's overall range. For example, MDC of the PEQ-MS in our study (0.55) was 13.7 percent of the scale range (4.0), and MDC of the 7-level response PEQ-MS used in the prior study (0.8) was 13.3 percent of the scale range (6.0) [49].

Instruments in our study that included more than 10 items (i.e., ABC, PEQ-MS, PLUS-M 12-item short form) generally had lower measurement error than instruments with fewer than 10 items (i.e., PROMIS-29 instruments, NQ-ACGC) or single-item instruments (i.e., PROMIS Pain Intensity and SCS). This may be expected, as the relationship between instrument length and measurement error is well established [57–58]. Estimates varied, however, even among items of similar length. For example, the PLUS-M 7-item short form has lower MDC (11.2% of the scale range) than the 8-item NQ-ACGC (17.0%). There was also variation in MDC estimates for the 4-item PROMIS instruments included in the PROMIS-29 Profile (17.5%–27.2%). All three versions of PLUS-M (i.e., computerized adaptive test, 12-item short form, and 7-item short form) had slightly lower estimates of MDC (9.1%–11.2% of the scale range) than the 16-item ABC (12.2%) and 12-item PEQ-MS (13.7%), which measure similar constructs. Thus, for measurement of mobility and balance in longitudinal applications (e.g., monitoring patients or participants over time), the PLUS-M 12-item short form and ABC are recommended.

Limitations

This study included a number of health status instruments designed to measure constructs of importance to prosthesis users, care providers, and researchers. PROMIS and Neuro-QoL instruments included in this study are available in lengths other than those tested. We used the 4-item versions of PROMIS instruments included in the PROMIS-29 Profile and the 8-item version of the NQ-ACGC. Estimates of reliability, MoA equivalence, measurement error, and detectable change derived here may therefore not apply to longer versions of these instruments (e.g., PROMIS Physical Function short forms are also available in 10- and 20-item lengths [59]). For example, evidence of potential floor and ceiling effects observed in this study may be a result of the administered versions' limited range of measurement. Although scores obtained with instruments from the same item bank are comparable, additional research is required to determine whether different lengths of these instruments function similarly in people with lower-limb loss [60].

The time between test and retest administrations in this study was relatively short (mean = 48.9 h). As such, respondents may have recalled responses to select questions. However, each survey included a large number of questions (n = 78; this is the sum of all items from all instruments) and participants were not allowed to take

retest surveys until 2 to 3 d had passed. Although test-retest periods of up to 2 wk may be advocated for self-report measures [11], evidence suggests that reliability of health status surveys is unaffected by test-retest periods of 2 d to 2 wk [61]. Thus, we believe it is unlikely that memory significantly affected results in the present study.

We evaluated MoA equivalence by comparing reliability estimates among three distinct administration modes (i.e., electronic, paper, and mixed). While significant differences would indicate lack of MoA equivalence, a more thorough evaluation of MoA would require multigroup confirmatory factor analysis (MGCFA) or, for measures developed within an item response theory framework, an assessment of differential item function (DIF). MGCFA may provide evidence of equivalent factor structures across MoAs, whereas DIF analyses may provide evidence of MoA equivalence at the item level. However, MGCFA and DIF analyses require significantly larger sample sizes than those in this study (i.e., 200–500 people in each arm) [24,62]. Given that the sample size in our study would likely bias results in favor of population invariance (i.e., MoA equivalence), we limited the scope of our evaluations to differences in estimates of reliability.

Only paper and electronic methods of administration modes were included in this study. Although we determined that most instruments were equivalent by MoA, results obtained here may not apply to other MoAs, such as face-to-face administration. Face-to-face administration introduces possible measurement biases (e.g., social desirability [63]) that may disproportionally affect responses relative to other MoAs. The setting of the interview may also have an effect on interview responses. Evidence to date regarding the equivalency of assisted (e.g., face-to-face interview) and self-report (e.g., paper or electronic survey) measures is limited [54], and further research in people with limb loss is required before equivalence can be verified.

Results of this study provide valuable evidence of test-retest reliability, MoA equivalency, measurement error, and detectable change for instruments suited to measuring prosthetic limb users. However, evidence of other measurement properties (e.g., validity and responsiveness) in this population may also guide how these instruments can and should be applied in clinical practice or research. Establishing evidence of these properties in people with lower-limb loss is beyond the scope of this study but may be considered a priority for future research.

CONCLUSIONS

The estimates of test-retest reliability, MoA equivalence, measurement error, and detectable change reported in this study can help clinicians and researchers better select, administer, and interpret outcomes from the self-report instruments. Reliability estimates showed that all of the studied measures are suited to group-level applications, and select instruments (i.e., ABC, PEQ-MS, and PLUS-M) are suited to individual-level applications based on thresholds established in the literature. Several instruments (i.e., PROMIS-29, NQ-ACGC, and SCS) showed evidence of potential floor or ceiling effects. SEM values derived in this study can allow users to calculate confidence intervals around individual scores. Similarly, the derived estimates for MDC can be used as a guide to assess whether differences in scores represent a true change or error in the measurement over repeated measurement.

ACKNOWLEDGMENTS

Author Contributions:
Study concept and design: B. J. Hafner, S. J. Morgan, R. L. Askew.
Analysis and interpretation of data: R. L. Askew, R. Salem, B. J. Hafner, S. J. Morgan.
Drafting of manuscript: B. J. Hafner, S. J. Morgan.
Critical revision of manuscript for important intellectual content: R. L. Askew, R. Salem.
Obtained funding: B. J. Hafner.
Financial Disclosures: The authors have declared that no competing interests exist.
Funding/Support: This material was based on work supported by the Orthotic and Prosthetic Education and Research Foundation (grant 2014-SGA-1), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant HD-065340), and the Department of Education (grant H133P080006).
Additional Contributions: The authors gratefully acknowledge Andre Kajlich, Meighan Rasley, and Olga Kildisheva for their assistance with participant recruitment and data collection.
Institutional Review: Approval of study procedures was obtained from a University of Washington Institutional Review Board. All study participants signed informed consent forms before study procedures were initiated.
Participant Follow-Up: The authors do not plan to notify participants of the publication of this study directly. This publication will be listed and linked on the research center's public Web site, which has been provided to all study participants.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the Orthotic and Prosthetic Education and Research Foundation, the National Institutes of Health, or the Department of Education.

REFERENCES

1. Heinemann AW, Connelly L, Ehrlich-Jones L, Fatone S. Outcome instruments for prosthetics: Clinical applications. Phys Med Rehabil Clin N Am. 2014;25(1):179–98. [PMID:24287247] http://dx.doi.org/10.1016/j.pmr.2013.09.002
2. Wedge FM, Braswell-Christy J, Brown CJ, Foley KT, Graham C, Shaw S. Factors influencing the use of outcome measures in physical therapy practice. Physiother Theory Pract. 2012;28(2):119–33. [PMID:21877943] http://dx.doi.org/10.3109/09593985.2011.578706
3. Jette DU, Halbert J, Iverson C, Miceli E, Shah P. Use of standardized outcome measures in physical therapist practice: Perceptions and applications. Phys Ther. 2009;89(2): 125–35. [PMID:19074618] http://dx.doi.org/10.2522/ptj.20080234
4. Gaunaurd I, Spaulding SE, Amtmann D, Salem R, Gailey R, Morgan SJ, Hafner BJ. Use of and confidence in administering outcome measures among clinical prosthetists: Results from a national survey and mixed-methods training program. Prosthet Orthot Int. 2015;39(4):314–21. [PMID:24827935] http://dx.doi.org/10.1177/0309364614532865
5. Stapleton T, McBrearty C. Use of standardised assessments and outcome measures among a sample of Irish occupational therapists working with adults with physical disabilities. Br J Occup Ther. 2009;72(2):55–64. http://dx.doi.org/10.1177/030802260907200203
6. Hatfield DR, Ogles BM. The use of outcome measures by psychologists in clinical practice. Prof Psychol Res Pr. 2004;35(5):485–91. http://dx.doi.org/10.1037/0735-7028.35.5.485
7. Sullivan JE, Crowner BE, Kluding PM, Nichols D, Rose DK, Yoshida R, Pinto Zipp G. Outcome measures for individuals with stroke: Process and recommendations from the American Physical Therapy Association neurology section task force. Phys Ther. 2013;93(10):1383–96. [PMID:23704035] http://dx.doi.org/10.2522/ptj.20120492
8. Potter K, Cohen ET, Allen DD, Bennett SE, Brandfass KG, Widener GL, Yorke AM. Outcome measures for individuals with multiple sclerosis: Recommendations from the American Physical Therapy Association Neurology Section task force. Phys Ther. 2014;94(5):593–608. [PMID:24363338] http://dx.doi.org/10.2522/ptj.20130149
9. Roach KE. Measurement of health outcomes: Reliability, validity and responsiveness. J Prosthet Orthot. 2006;18(1S): 8–12. http://dx.doi.org/10.1097/00008526-200601001-00003
10. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, Schwartz C, Revicki DA, Moinpour CM, McLeod LD, Lyons JC, Lenderking WR, Hinds PS, Hays RD, Greenhalgh J, Gershon R, Feeny D, Fayers PM, Cella D, Brundage M, Ahmed S, Aaronson NK, Butt Z. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22(8):1889–1905. [PMID:23288613] http://dx.doi.org/10.1007/s11136-012-0344-y
11. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. [PMID:17161752] http://dx.doi.org/10.1016/j.jclinepi.2006.03.012
12. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health. 2007;10(Suppl 2):S94–105. [PMID:17995479] http://dx.doi.org/10.1111/j.1524-4733.2007.00272.x
13. Lohr KN. Rating the strength of scientific evidence: Relevance for quality improvement programs. Int J Qual Health Care. 2004;16(1):9–18. [PMID:15020556] http://dx.doi.org/10.1093/intqhc/mzh005
14. Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, Rothman M. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res. 2000; 9(8):887–900. [PMID:11284208] http://dx.doi.org/10.1023/A:1008996223999
15. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998;2(14):i–iv, 1–74. [PMID:9812244]
16. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med. 2000;30(1):1–15. [PMID:10907753] http://dx.doi.org/10.2165/00007256-200030010-00001
17. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York (NY): McGraw-Hill; 1994.
18. Hood K, Robling M, Ingledew D, Gillespie D, Greene G, Ivins R, Russell I, Sayers A, Shaw C, Williams J. Mode of data elicitation, acquisition and response to surveys: A systematic review. Health Technol Assess. 2012;16(27):1–162. [PMID:22640750] http://dx.doi.org/10.3310/hta16270

19. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. Value Health. 2008;11(2):322–33. [PMID:18380645] http://dx.doi.org/10.1111/j.1524-4733.2007.00231.x
20. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. J Clin Epidemiol. 2004;57(10):1008–18. [PMID:15528051] http://dx.doi.org/10.1016/j.jclinepi.2004.02.007
21. Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical change of significance: Issues and methods. Am J Occup Ther. 1988;42(3):156–63. [PMID:2451425] http://dx.doi.org/10.5014/ajot.42.3.156
22. Condie E, Scott H, Treweek S. Lower limb prosthetic outcome measures: A review of the literature 1995 to 2005. J Prosthet Orthot. 2006;18(1S):13–45. http://dx.doi.org/10.1097/00008526-200601001-00004
23. Suresh K. An overview of randomization techniques: An unbiased assessment of outcome in clinical research. J Hum Reprod Sci. 2011;4(1):8–11. [PMID:21772732] http://dx.doi.org/10.4103/0974-1208.82352
24. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, Lenderking WR, Cella D, Basch E; ISPOR ePRO Task Force. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value Health. 2009;12(4):419–29. [PMID:19900250] http://dx.doi.org/10.1111/j.1524-4733.2008.00470.x
25. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101–10. [PMID:9463853] http://dx.doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E
26. Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment center. Qual Life Res. 2010;19(5):677–85. [PMID:20306332] http://dx.doi.org/10.1007/s11136-010-9634-4
27. Paulsen A, Overgaard S, Lauritsen JM. Quality of data entry using single entry, double entry and automated forms processing—An example based on a study of patient-reported outcomes. PLoS ONE. 2012;7(4):e35087. [PMID:22493733] http://dx.doi.org/10.1371/journal.pone.0035087
28. Hafner BJ, Spaulding SE, Salem R, Morgan SJ, Gaunaurd I, Gailey R. Prosthetists' perceptions and use of outcome measures in clinical practice: Long-term effects of focused continuing education. Prosthet Orthot Int. 2016. Epub ahead of print. [PMID:27638912] http://dx.doi.org/10.1177/0309364616664152
29. Amtmann D, Abrahamson D, Morgan S, Salem R, Askew R, Gailey R, Gaunaurd I, Kajlich A, Hafner B. The PLUS-M: Item bank of mobility for prosthetic limb users. Qual Life Res. 2014;23(1S):39–40. http://dx.doi.org/10.1007/s11136-014-0769-6
30. Morgan SJ, Amtmann D, Abrahamson DC, Kajlich AJ, Hafner BJ. Use of cognitive interviews in the development of the PLUS-M item bank. Qual Life Res. 2014;23(6):1767–75. [PMID:24442531] http://dx.doi.org/10.1007/s11136-013-0618-z
31. Legro MW, Reiber GD, Smith DG, del Aguila M, Larsen J, Boone D. Prosthesis evaluation questionnaire for persons with lower limb amputations: Assessing prosthesis-related quality of life. Arch Phys Med Rehabil. 1998;79(8):931–8. [PMID:9710165] http://dx.doi.org/10.1016/S0003-9993(98)90090-9
32. Powell LE, Myers AM. The Activities-specific Balance Confidence (ABC) Scale. J Gerontol A Biol Sci Med Sci. 1995;50A(1):M28–34. [PMID:7814786] http://dx.doi.org/10.1093/gerona/50A.1.M28
33. Franchignoni F, Giordano A, Ferriero G, Orlandini D, Amoresano A, Perucca L. Measuring mobility in people with lower limb amputation: Rasch analysis of the mobility section of the prosthesis evaluation questionnaire. J Rehabil Med. 2007;39(2):138–44. [PMID:17351696] http://dx.doi.org/10.2340/16501977-0033
34. Sakakibara BM, Miller WC, Backman CL. Rasch analyses of the Activities-specific Balance Confidence Scale with individuals 50 years and older with lower-limb amputations. Arch Phys Med Rehabil. 2011;92(8):1257–63. [PMID:21704978] http://dx.doi.org/10.1016/j.apmr.2011.03.013
35. Cella D, Nowinski C, Peterman A, Victorson D, Miller D, Lai JS, Moy C. The neurology quality-of-life measurement initiative. Arch Phys Med Rehabil. 2011;92(10 Suppl):S28–36. [PMID:21958920] http://dx.doi.org/10.1016/j.apmr.2011.01.025
36. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R; PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010;63(11):1179–94. [PMID:20685078] http://dx.doi.org/10.1016/j.jclinepi.2010.04.011
37. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M; PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement

Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. Med Care. 2007;45(5 Suppl 1):S3–11. [PMID:17443116] http://dx.doi.org/10.1097/01.mlr.0000258615.42478.55

38. Hanspal RS, Fisher K, Nieveen R. Prosthetic socket fit comfort score. Disabil Rehabil. 2003;25(22):1278–80. [PMID:14617445] http://dx.doi.org/10.1080/09638280310001603983
39. Cohen RJ, Swerdlik ME, Phillips SM. Psychological testing and assessment: An introduction to tests and measurement. Mountain View (CA): Mayfield Publishing Co; 1996.
40. D'Agostino RB, Belanger A, D'Agostino RB Jr. A suggestion for using powerful and informative tests of normality. Am Stat. 1990;44(4):316–21. http://dx.doi.org/10.2307/2684359
41. Royston JP. sg3.5: Comment on sg3.4 and an improved D'Agostino test. Stata Tech Bull. 1991;3:23–4.
42. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8. [PMID:18839484] http://dx.doi.org/10.1037/0033-2909.86.2.420
43. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231–40. [PMID:15705040]
44. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30–46. http://dx.doi.org/10.1037/1082-989X.1.1.30
45. Feldt LS, Woodruff DJ, Salih FA. Statistical inference for coefficient alpha. Appl Psychol Meas. 1987;11(1):93–103. http://dx.doi.org/10.1177/014662168701100107
46. Kraemer HC. Extension of Feldt's approach to testing homogeneity of coefficients of reliability. Psychometrika. 1981;46(1):41–5. http://dx.doi.org/10.1007/BF02293917
47. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita. Florence (Italy): Libreria Internazionale Seeber; 1936.
48. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol. 2011; 64(5):507–16. [PMID:21447427] http://dx.doi.org/10.1016/j.jclinepi.2010.11.018
49. Resnik L, Borgia M. Reliability of outcome measures for people with lower-limb amputations: Distinguishing true change from statistical error. Phys Ther. 2011;91(4):555–65. [PMID:21310896] http://dx.doi.org/10.2522/ptj.20100287
50. de Laat FA, Rommers GM, Geertzen JH, Roorda LD. Construct validity and test-retest reliability of the questionnaire rising and sitting down in lower-limb amputees. Arch Phys Med Rehabil. 2011;92(8):1305–10. [PMID:21807151] http://dx.doi.org/10.1016/j.apmr.2011.03.016
51. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651–7. [PMID:21732199] http://dx.doi.org/10.1007/s11136-011-9960-1
52. Pezzin LE, Dillingham TR, Mackenzie EJ, Ephraim P, Rossbach P. Use and satisfaction with prosthetic limb devices and related services. Arch Phys Med Rehabil. 2004;85(5):723–9. [PMID:15129395] http://dx.doi.org/10.1016/j.apmr.2003.06.002
53. Miller WC, Deathe AB, Speechley M. Psychometric properties of the Activities-specific Balance Confidence Scale among individuals with a lower-limb amputation. Arch Phys Med Rehabil. 2003;84(5):656–61. [PMID:12736877] http://dx.doi.org/10.1016/S0003-9993(02)04807-4
54. Rutherford C, Costa D, Mercieca-Bebber R, Rice H, Gabb L, King M. Mode of administration does not cause bias in patient-reported outcome results: A meta-analysis. Qual Life Res. 2016;25(3):559–74. [PMID:26334842] http://dx.doi.org/10.1007/s11136-015-1110-8
55. Muehlhausen W, Doll H, Quadri N, Fordham B, O'Donohoe P, Dogar N, Wild DJ. Equivalence of electronic and paper administration of patient-reported outcome measures: A systematic review and meta-analysis of studies conducted between 2007 and 2013. Health Qual Life Outcomes. 2015;13(1):167. [PMID:26446159]
56. Toepoel V, Das M, van Soest A. Design of web questionnaires: The effect of layout in rating scales. J Off Stat. 2009;25(4):509–28. http://dx.doi.org/10.2139/ssrn.903740
57. Lord FM. Tests of the same length do have the same standard error of measurement. Educ Psychol Meas. 1959; 19(2):233–9. http://dx.doi.org/10.1177/001316445901900208
58. Gardner PL. Test length and the standard error of measurement. J Educ Meas. 1970;7(4):271–3. http://dx.doi.org/10.1111/j.1745-3984.1970.tb00728.x
59. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: Assessment of physical function. J Rheumatol. 2014;41(1):153–8. [PMID:24241485] http://dx.doi.org/10.3899/jrheum.130813
60. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007;16(Suppl 1):133–41. [PMID:17401637] http://dx.doi.org/10.1007/s11136-007-9204-6
61. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol. 2003;56(8):

730–5. [PMID:12954464]
http://dx.doi.org/10.1016/S0895-4356(03)00084-2

62. Comrey AL, Lee HB. A first course in factor analysis. 2nd ed. Hoboken (NJ): Taylor and Francis; 2013.
63. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. J Appl Psychol. 2003;88(5):879–903. [PMID:14516251]
http://dx.doi.org/10.1037/0021-9010.88.5.879

Submitted for publication December 6, 2015. Accepted in revised form March 29, 2016.

This article and any supplementary material should be cited as follows:
Hafner BJ, Morgan SJ, Askew RL, Salem R. Psychometric evaluation of self-report outcome measures for prosthetic applications. J Rehabil Res Dev. 2016;53(6):797–812.
http://dx.doi.org/10.1682/JRRD.2015.12.0228